Alignment-based protein mutational landscape prediction: doing more with less

Marina Abakarova[1,2+], Céline Marquet[3,4+], Michael Rera[2], Burkhard Rost[3,5,6], Elodie Laine[1*]

December 13, 2022

[1] Sorbonne Université, CNRS, IBPS, Laboratory of Computational and Quantitative Biology (LCQB), UMR 7238, Paris, 75005, France
[2] Université Paris Cité, INSERM UMR U1284, 75004 Paris, France
[3] Department of Informatics, Bioinformatics and Computational Biology - i12, TUM-Technical University of Munich, Boltzmannstr. 3, Garching, 85748 Munich, Germany
[4] TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Boltzmannstr. 11, 85748 Garching, Germany
[5] Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, Garching, 85748 Munich, Germany
[6] TUM School of Life Sciences Weihenstephan (TUM-WZW), Alte Akademie 8, Freising, Germany

[+] equally contributing authors
[*] corresponding author: elodie.laine@sorbonne-universite.fr

Abstract

Recent efforts for democratising protein structure prediction have leveraged the MMseqs2 algorithm to efficiently generate multiple sequence alignments with high diversity and a limited number of sequences. Here, we investigated the usefulness of this strategy for mutational outcome prediction. We place ourselves in a context where we only exploit information coming from the input alignment for making predictions. Through a large-scale assessment of ∼1.5M missense variants across 72 protein families, we show that the MMseqs2-based protocol implemented in ColabFold compares favourably with tools and resources relying on profile-Hidden Markov Models. Our study demonstrates the feasibility of simultaneously providing high-quality and compute-efficient alignment-based predictions for the mutational landscape of entire proteomes.

1 Introduction

In recent years, tremendous progress has been achieved in the prediction of protein 3D structures and mutational landscapes [1, 2] by leveraging the wealth of publicly available natural protein sequence data [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. State-of-the-art predictors capture arbitrary range dependencies between amino acid residues by implicitly accounting for global sequence contexts or explicitly exploiting structured information coming from alignments of evolutionary related protein sequences. Very efficient algorithms, *e.g.* MMseqs2 [14], allow for identifying homologous sequences and aligning them on a mass scale. Others relying on profile hidden Markov models (HMMs), such as JackHMMer/HMMer [15], carefully generate very large families, achieving a very high sensitivity. Several large-scale resources like Pfam [16] and ProteinNet [17] give access to pre-computed multiple sequence alignments (MSAs) built from profile HMMs. These MSAs are associated with curated protein families in Pfam, or with experimentally resolved protein 3D structures in ProteinNet. The depth, quality, and computational cost of a MSA are important factors contributing to its effective usefulness. Nevertheless, precisely assessing the impact of expanding or filtering out sequences on predictive performance is difficult. For protein structure prediction, Mirdita and co-authors showed that AlphaFold2 original performance could be attained with much smaller and cheaper alignments through the MMseqs2 [14]-based strategy implemented in ColabFold [3].

In this work, we tested whether the same gain could be achieved for mutational outcome prediction. We compared the prediction accuracy achieved by Global Epistatic Model for predicting Mutational Effects (GEMME) [20] from MSAs generated using the ColabFold's MMseqs2-based protocol [3, 14] versus three classical workflows relying on profile HMMs [17, 16, 18]. GEMME is a fast MSA-based mutational outcome predictor relying on a few biologically meaningful and interpretable parameters. It performs on-par with statistical inference-based methods estimating pairwise couplings [21] and also deep learning-based methods, including family-specific models [22, 23, 24, 25] as well as high-capacity protein language models trained across protein families [18, 26, 27] (**Fig. S1**, see also [24, 26, 20] for quantitative comparisons). We assessed GEMME predictions against a large collection of 87 Deep Mutational Scanning experiments (DMS) totalling ∼1.5M missense variants across 72 diverse protein families [18]. We used the Spearman rank correlation coefficient to quantify the accuracy of the predictions, as previously done by us and others [18, 27, 20].

2 Materials and Methods

2.1 DMS benchmark set

We downloaded the ProteinGym substitution benchmark [18] from the following repository: `https://github.com/OATML-Markslab/Tranception`. It contains measurements from 87 DMS collected for 72 proteins of various sizes (between 72 and 3,423 residue long), functions (*e.g.* kinases, ion channels, g-protein coupled receptors, polymerases, transcription factors, tumor suppressors), and origins (**Fig. S2A-C**). The DMS cover a wide range of functional properties, including thermostability, ligand binding, aggregation, viral replication, and drug resistance. Up to four experiments are reported for each protein (**Fig. S2D**). Although the benchmark mostly focuses on single point mutations, it also reports multiple amino-acid variant measurements for 11 proteins (**Table S1**).

2.2 MSA resources and protocols

We considered four different MSA generation protocols and resources, referred to as ProteinGym, ColabFold, ProteinNet and Pfam (**Table 1**). They represent a variety of choices in terms of sequence database, search algorithm and sequence context. Two protocols, ColabFold and ProteinGym, were available for all 87 DMS (from 72 proteins) from the ProteinGym benchmark. ProteinNet was available for 51 (from 42 proteins), Pfam for 52 (from 39 proteins). When comparing two methods, we reduced the Spearman rank calculations to their common positions.

Table 1: **Details about the MSA generation protocols.**

Name	Databases	Search algorithms	Fine tuning	#(covered proteins)[a]	#(sequences) Min - Max
ProteinGym	UniRef100 [12]	JackHMMer [15]	yes[b]	72	44 - 539,868
ColabFold	UniRef30 [12] and ColabFold env.[c] [3]	MMseqs2 [14]	no	72	126 - 24,269
ProteinNet	UniParc[d] [31] and IMG [13]	JackHMMer [15]	no	42	249 - 1,389,216
Pfam	UniProtKB [5]	HMMer [15]	yes[e]	39[f]	134 - 283,380

[a]We indicate the number of proteins treated with each protocol, out of the 72 proteins comprised in the ProteinGym substitution benchmark. [b]For each protein, 9 MSAs were generated by exploring bit score thresholds from 0.1 to 0.9 and the MSA leading to the highest number of significant Evolutionary Couplings [21] was retained. [c]ColabFold environmental database contains BFD [6], which includes UniProt/TrEMBL+Swissprot, Mgnify [9], MetaEuk [10], SMAG [4], TOPAZ [34], MGV [7], GPD [8], and MetaClust2 [11]. [d] UniParc, for UniProt Archive, is a non-redundant archive of protein sequences extracted from more than 10 public databases, including UniProtKB, Ensembl [35], PDB, FlyBase [36] and WormBase [37]. [e]For each Pfam family, the profile HMM used to query UniProtKB was hand curated, and the score threshold used to select the sequences was set manually. [f]For this protocol, we considered a non-redundant subset of 59 proteins.

The ColabFold protocol [3] relies on the very fast MMseqs2 method [14] (3 iterations) to search against UniRef30, a 30% sequence identity clustered database based on UniProt [5], and a novel database

compiling several environmental sequence sets (**Table 1**). It maximises diversity while limiting the number of sequences through an expand-and-filter strategy. Specifically, it iteratively identifies representative hits, expand them with their cluster members, and filters the latter before adding them to the MSA. We used the same sequence queries as those defined in ProteinGym. For all but 5 proteins, the query corresponds to the full-length UniProt sequence. For each query, we generated two MSAs by searching against UniRef30 and ColabFold environmental database, respectively, and we then concatenated them.

The ProteinGym protocol [18] relies on the highly sensitive homology detection method JackHMMer [15] (5 iterations) to search against UniRef100 [12], the non-redundant version of UniProt (**Table 1**). JackHMMer is part of the HMMer suite and is based on profile hidden Markov models (HMMs). This protocol is relatively costly, with up to several hours for a single input MSA. The MSAs generated with this protocol have been widely used to assess mutational outcome predictors [18, 21]. In this work, we took the alignments provided with the ProteinGym benchmark [18].

The ProteinNet protocol [17] also performs 5 iterations of JackHMMER, but it extends the sequence database to the whole UniProt Archive (Uniparc) [31] complemented with metagenomic sequences from IMG [13] (**Table 1**). Another difference from ProteinGym is that the queries correspond to sequences extracted from experimentally determined protein structures available in the PDB [19]. The MSAs are readily available and organised in a series of data sets, each one encompassing all proteins structurally characterised prior to different editions of the Critical Assessment of protein Structure Prediction (CASP) [32]. We chose the most complete set, namely ProteinNet12. It covers all proteins whose structure was deposited in the PDB before 2016, the year of CASP round XII [33]. For each protein from the ProteinGym benchmark, we retrieved the corresponding PDB codes from the Uniprot website (`https://www.uniprot.org`) and picked up the structure with the highest coverage among those represented in ProteinNet12 (**Table S1**). We could treat 42 proteins, out of 72 in total. For the remaining ones, the positions covered by the available MSAs were out of the range of mutated positions.

The Pfam database [16] is a resource of manually curated protein domain families. Each family, sometimes referred to as a Pfam-A entry, is associated with a profile HMM built using a small number of representative sequences, and several MSAs. We chose to work with the full UniProt alignment, obtained by searching the family-specific profile-HMM against UniProtKB (**Table 1**). The proteins sharing the same domain composition will have exactly the same MSAs. To avoid such redundancy, we focused on a subset of 59 proteins extracted with an adjusted version of UniqueProt [29, 30]. Instead of PSI-BLAST we used MMseqs2 to improve runtime, and discarded alignments of less than 50 residues for pairs of sequences with at least 180 residues to prevent very short alignments from removing longer sequences. For each protein, we first retrieved its Pfam domain composition and downloaded the corresponding MSAs from the Pfam website (`https://pfam.xfam.org`, release 34.0). We could retrieve at least one (and up to 5) MSA overlapping with the range of mutated positions for 39 proteins (**Table S1**). Each detected Pfam domain appears only once in the set.

3 Results and Discussion

3.1 The ColabFold protocol leads to the most accurate predictions

ColabFold and ProteinGym are the best performing protocols and the only ones covering all ∼1.5M mutations from the ProteinGym benchmark (**Table 2**). The ColabFold protocol allows obtaining more accurate predictions for two thirds of the DMS (**Fig. 1A**), while producing MSAs with substantially fewer sequences (**Fig. S3**). More precisely, for the proteins with abundant sequence information (**Table 2**, "high" category based on ProteinGym MSAs), the accuracy is higher by $\Delta\bar{\rho} = 0.032$ on average and the MSAs are shallower (**Fig. 1B**, N_{eff} ratio < 1, see red triangles). In fact, all proteins falling in the "high" alignment depth category ($N_{eff}/L > 100$, see *Materials and Methods*) based on their ProteinGym MSAs would be reclassified in the "medium" category ($1 < N_{eff}/L < 100$) based on their ColabFold MSAs (**Fig. S4**). This observation highlights the relevance of ColabFold's MMseqs2-based expand-and-filter strategy for these cases. For the "medium" and "low" categories, the results are less clear. On the one hand, the ColabFold protocol increases the alignment depth for 24 proteins belonging to these categories (**Fig. 1B**, see green and blue triangles with N_{eff} ratio > 1). For instance, for the SARS-CoV-2 Replicase polyprotein 1ab, GEMME could make predictions only with the ColabFold MSA, the variability of the ProteinGym MSA being too low (**Fig. 1A**, see null x-value). Overall, the accuracy gain resulting from the increased MSA depth is limited ($\Delta\bar{\rho} = 0.015 \pm 0.045$). On the other hand, ColabFold produces very shallow MSAs for the polymerases PA and PB2 from influenza A virus (UniProt names: PA_I34A1 and A4D664_9INFA, respectively), 20 times shallower than those produced by ProteinGym, resulting in a dramatic deterioration of the prediction accuracy for these proteins (**Fig. 1B**, see the two outliers, $\Delta\rho \sim -0.3$). This behaviour does not extend to the other viral proteins from the benchmark.

3.2 Expanding the sequence search space marginally improves prediction accuracy

The ColabFold MSAs result from applying an MMseqs2-based search, expand and filter algorithm to both the UniRef30 database, and the ColabFold database comprising UniProt/TrEMBL, Swissprot, and several collections of environmental sequences (**Table 1**). We found that the ColabFold database marginally contributed to the mutational outcome predictions (**Fig. S5**). It proved necessary in only one case, the human SC6A4. In addition, it slightly improved prediction accuracy for a few viral proteins, yet without allowing reaching a good agreement with the experimental measurements – the Spearman rank correlation remains below 0.3 (**Fig. S5**). By contrast, it significantly deteriorated the predictions for the human KCNH2 by $\Delta\rho = -0.14$. The limited influence of metagenomics can also be observed when using JackHMMer as the search algorithm, as attested by the similar performance obtained for ProteinGym (UniRef100) and ProteinNet (UniParc and IMG, see **Table 2**). By looking at the per-DMS Spearman rank correlations (**Fig. 2A**), we could identify a few human proteins, namely P53, BRCA1, SUMO1, and YAP1, as well as IF1 and CCDB from *E. coli*, that benefited from the additional information exploited by ProteinNet. By contrast, the Spearman rank correlation computed for the yeast protein GAL4 dropped dramatically, from 0.497 to 0.217. This result illustrates the interest of considering the full sequence context. While the ProteinGym protocol could retrieve 16,159 sequences by querying the full-length protein sequence, the ProteinNet MSA,

Table 2: **Average Spearman's rank correlation between predicted values and experimental measurements on the ProteinGym substitution benchmark.**

Set	Class	#(proteins)	#(DMS)	ColabFold	ProteinGym	ProteinNet	Pfam
All		72	87	**0.470**	0.463	-	-
	Low	14	20	**0.453**	0.444	-	-
	Medium	43	17	0.443	**0.446**	-	-
	High	15	50	**0.552**	0.520	-	-
	Human	26	32	**0.445**	0.436	-	-
	Eukaryote	10	13	**0.500**	0.479	-	-
	Prokaryote	17	21	**0.529**	0.505	-	-
	Virus	19	21	0.429	**0.451**	-	-
ProteinNet		42	51	**0.507**	0.497	0.495	-
	Human	19	23	**0.484**	0.466	0.477	-
	Eukaryote	6	7	**0.539**	0.531	0.495	-
	Prokaryote	13	17	**0.562**	0.536	0.540	-
	Virus	4	4	0.353	**0.453**	0.410	-
Pfam		39	52	**0.463**	0.440	-	0.432
	Human	15	20	**0.440**	0.423	-	0.407
	Eukaryote	7	10	**0.462**	0.448	-	0.436
	Prokaryote	9	13	**0.517**	0.489	-	0.496
	Virus	8	9	**0.438**	0.399	-	0.391

The N_{eff} categories *Low*, *Medium* and *High* were taken from [18] and correspond to the ProteinGym alignments. We use this classification as a reference, although proteins may change category between the different protocols (see **Fig. S4**). The Spearman rank correlations are computed either over all residues from the target sequences, or only the residue ranges covered by ProteinNet and Pfam, respectively. The correlations over the full-length versus partial proteins are comparable for ColabFold and ProteinGym protocols (**Fig. S6**).

which covers a very small portion of the protein (**Fig. 2A**, 6% that is 55 residues out of 881, PDB code: 1HBW), comprises only 249 sequences.

3.3 A domain-focused perspective

The residue spans defined by the Pfam and ProteinNet MSAs correspond to well-curated or well-folded protein domains. One may wonder whether the predictions are better in these regions compared to unannotated or disordered regions. In our experiment, we did not observe such a trend. The ColabFold and ProteinGym MSAs yielded comparable Spearman correlation coefficients over the full-length protein and over the regions annotated as Pfam domains or with experimentally resolved 3D structures (**Fig. S6**). Moreover, reconstructing a protein's mutational landscape by combining predictions coming from different MSAs, each one representing a curated Pfam domain, proved less accurate than building a single query-specific full-length MSA (**Fig. S7**). Indeed, the ColabFold strategy led to a higher Spearman rank correlation than the Pfam protocol for 70% of the considered DMS (**Fig. 2C**). For the remaining 30%, the gain brought by Pfam does not exceed $\Delta\rho_{max} = 0.077$.

4 Conclusion

Overall, this study identified ColabFold as the best suited MSA generation protocol for assessing protein mutational outcomes. It yields the best performance and allows covering protein regions lacking structural data or domain annotations. It limits the number of sequences, thus preventing memory issues. It is faster than classical homology detection methods by orders of magnitude. The study also showed that the alignment depth is not a good indicator of the prediction accuracy as one might expect. The Spearman rank correlation can be as good as 0.7 even with shallow alignments. And above a certain threshold, adding more sequences does not improve the predictions. Moreover, extending the sequence search space to environmental datasets only marginally improves the accuracy of the predictions. Finally, readily available resources such as ProteinNet and Pfam are valid options, but they only provide a partial coverage of the query proteins. This study demonstrates the feasibility of MSA-based computational scans of entire proteomes at a very large scale. Combining ColabFold with GEMME, it takes only a few days to generate the complete single-mutational landscape of the human proteome on the supercomputer “MeSU” of Sorbonne University (64 CPUs from Intel Xeon E5-4650L processors, 910GB shared RAM memory).

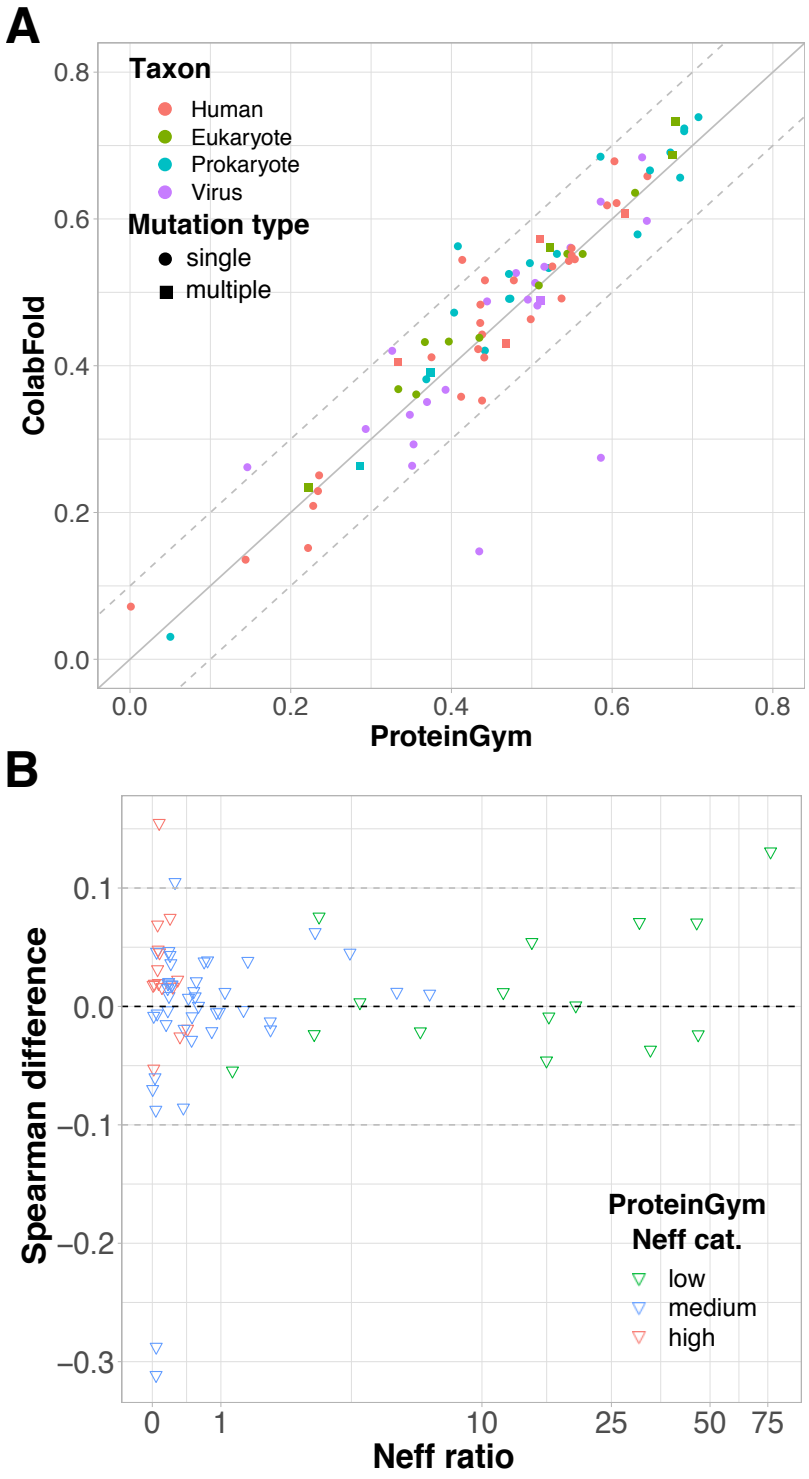

Figure 1: **Comparison of the ProteinGym and ColabFold protocols. A.** GEMME's Spearman rank correlation coefficients (ρ) computed against the 87 DMS sets from the ProteinGym substitution benchmark. The input MSAs were generated using the ProteinGym (x-axis) or ColabFold (y-axis) protocols. The colors indicate the taxons of the target sequences and the shapes indicate whether the experiment contains only single mutations (circle) or also multiple mutations (square). **B.** Differences in ρ values in function of the number of effective sequence (N_{eff}) ratio. Positive values correspond to ColabFold performing better than ProteinGym. Each point (triangle) corresponds to a given input MSA (*i.e.* a given target sequence) and its y-value is averaged over the set of DMS experiments (between 1 and 4, see **Fig. S2**) associated to it. The colors indicate the depth of the ProteinGym MSAs, either low, medium or high, as defined in [18] (see also *Materials and Methods*).

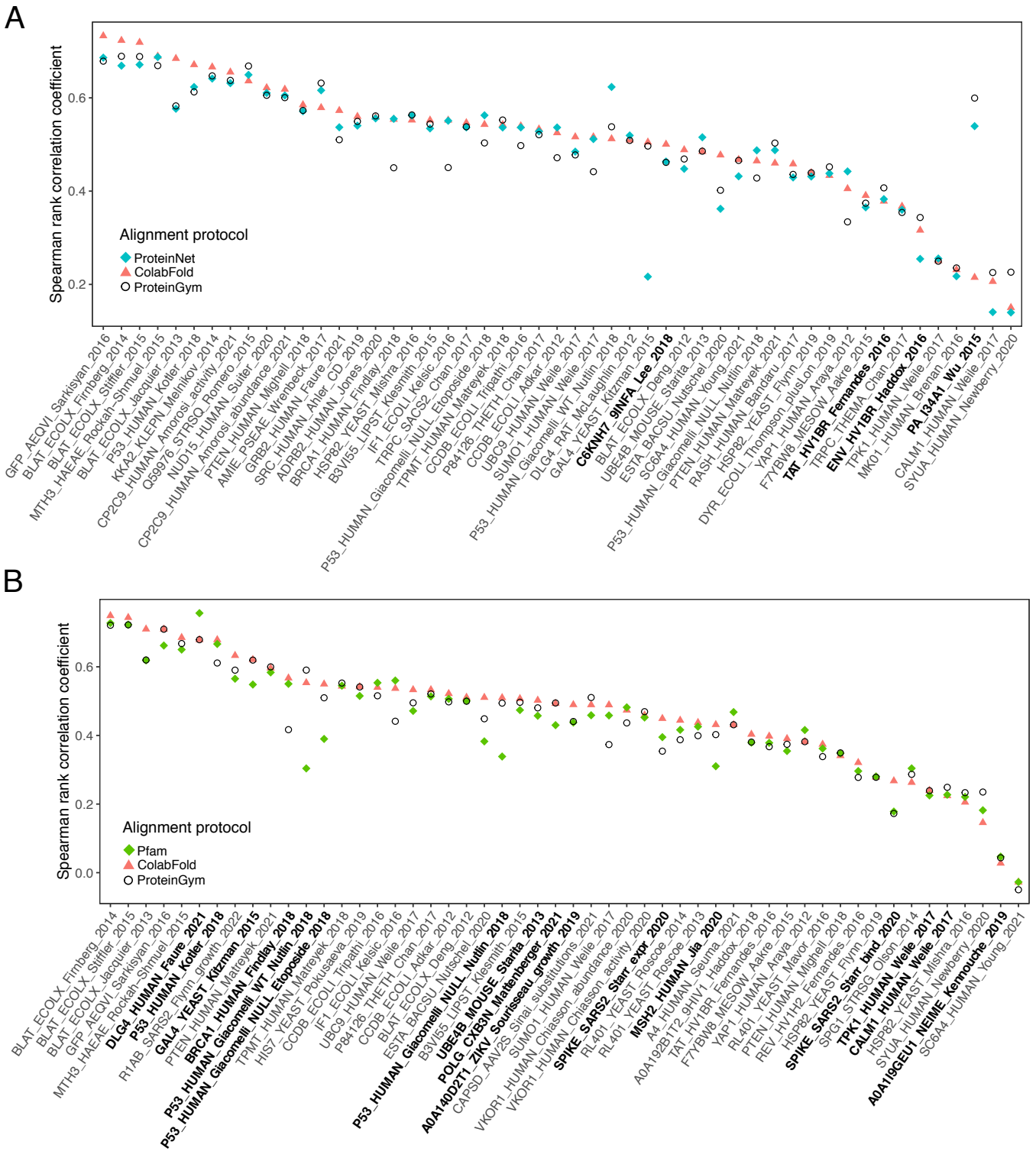

Figure 2: **Performance comparison between the different MSA generation protocols. A.** Comparison of ProteinNet, ColabFold and ProteinGym, focusing on the 51 DMS covered by ProteinNet (x-axis). The Spearman rank correlation coefficients are computed over the residue spans covered by ProteinNet MSAs for all methods. The DMS associated to viral proteins are highlighted in bold. **B.** Comparison of Pfam, ColabFold and ProteinGym, focusing on the 52 DMS covered by Pfam (x-axis). The Spearman rank correlation coefficients are computed over the residue spans covered by Pfam MSAs for all methods. The DMS associated to proteins containing more than one Pfam domains are highlighted in bold on the x-axis.

References

1. Method of the Year 2021: Protein structure prediction. Nature Methods. 2022;19(1):1–1.
2. Laine E, Eismann S, Elofsson A, Grudinin S. Protein sequence-to-structure learning: Is this the end (-to-end revolution)? Proteins: Structure, Function, and Bioinformatics. 2021;89(12):1770–1786.
3. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nature Methods. 2022;p. 1–4.
4. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics. 2022;2(5):100123.
5. UniProt: the universal protein knowledgebase in 2021. Nucleic acids research. 2021;49(D1):D480–D489.
6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.
7. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nature microbiology. 2021;6(7):960–970.
8. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098–1109.
9. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic acids research. 2020;48(D1):D570–D578.
10. Levy Karin E, Mirdita M, Söding J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 2020;8(1):1–15.
11. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nature communications. 2018;9(1):1–8.
12. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–932.
13. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research. 2014;42(D1):D26–D31.
14. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology. 2017;35(11):1026–1028.
15. Eddy SR. Accelerated profile HMM searches. PLoS computational biology. 2011;7(10):e1002195.
16. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic acids research. 2021;49(D1):D412–D419.

17. AlQuraishi M. ProteinNet: a standardized data set for machine learning of protein structure. BMC bioinformatics. 2019;20(1):1–10.

18. Notin P, Dias M, Frazer J, Hurtado JM, Gomez AN, Marks D, et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In: International Conference on Machine Learning. PMLR; 2022. p. 16990–17017.

19. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The Protein Data Bank. Acta Crystallographica Section D: Biological Crystallography. 2002 Jun;58(6):899–907. Available from: `http://scripts.iucr.org/cgi-bin/paper?an0594`.

20. Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. Molecular biology and evolution. 2019;36(11):2604–2619.

21. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nature biotechnology. 2017;35(2):128–135.

22. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–95.

23. Shin JE, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, et al. Protein design and variant prediction using autoregressive generative models. Nature communications. 2021;12(1):1–11.

24. Trinquier J, Uguzzoni G, Pagnani A, Zamponi F, Weigt M. Efficient generative modeling of protein sequences using simple autoregressive models. Nature communications. 2021;12(1):1–11.

25. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature methods. 2018;15(10):816–822.

26. Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. Human genetics. 2021;p. 1–19.

27. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems. 2021;34:29287–29303.

28. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. In: International Conference on Learning Representations; 2020. .

29. Olenyi T, Bernhofer M, Miridita M, Steinegger M, Rost B. Rostclust redundancy reduction. Manuscript in preparation. 2022;Department of Informatics, Technical University of Munich.

30. Mika S, Rost B. UniqueProt: creating representative protein sequence sets. Nucleic acids research. 2003;31(13):3789–3791.

31. Consortium U, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2018;46(5):2699.

242 32. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of pro-
243 tein structure prediction (CASP)—Round XIV. Proteins: Structure, Function, and Bioinformatics.
244 2021;89(12):1607–1617.

245 33. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of
246 protein structure prediction (CASP)—Round XII. Proteins: Structure, Function, and Bioinformatics.
247 2018;86:7–15.

248 34. Alexander H, Hu SK, Krinos AI, Pachiadaki M, Tully BJ, Neely CJ, et al. Eukaryotic genomes from a
249 global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. bioRxiv.
250 2022;p. 2021–07.

251 35. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic
252 acids research. 2021;49(D1):D884–D891.

253 36. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0:
254 the next generation. Nucleic acids research. 2019;47(D1):D759–D765.

255 37. Davis P, Zarowiecki M, Arnaboldi V, Becerra A, Cain S, Chan J, et al. WormBase in 2022—data,
256 processes, and tools for analyzing Caenorhabditis elegans. Genetics. 2022;220(4):iyac003.

257 38. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using
258 pseudolikelihoods to infer Potts models. Physical Review E. 2013;87(1):012707.

259 39. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote
260 homology detection and deep protein annotation. BMC bioinformatics. 2019;20(1):1–15.

Supplementary methods

Alignment depth

We measured the alignment depth as the ratio of the effective number of sequences N_{eff} by the number of positions L. The effective number of sequences is computed as a sum of weights [38],

$$N_{eff} = \sum_{s}^{N} \pi_s, \tag{1}$$

where N is the number of sequences in the MSA and π_s is the weight assigned to sequence $\mathbf{x}^{(s)}$, computed as

$$\pi_s = (\sum_{t}^{N} I[D_H(\mathbf{x}^{(s)}, \mathbf{x}^{(t)}) < \theta_{ID}])^{-1}, \tag{2}$$

where $D_H(\mathbf{x}^{(s)}, \mathbf{x}^{(t)})$ is the normalised Hamming distance between the sequences $\mathbf{x}^{(s)}$ and $\mathbf{x}^{(t)}$ and θ_{ID} is a predefined neighbourhood size (percent divergence). Hence, the weight of a given sequence reflects how dissimilar it is to the other sequences in the MSA. To be consistent with [18], we set $\theta_{ID} = 0.2$ (80% sequence identity) for eukaryotic and prokaryotic proteins, and $\theta_{ID} = 0.01$ (99% sequence identity) for viral proteins.

In [18], MSAs are labeled as Low, Medium or High depending on the ratio N_{eff}/L_{cov}, where L_{cov} is the number of positions with less than 30% gaps. Specifically, MSAs with $N_{eff}/L_{cov} < 1$ are considered as shallow ('Low' group) whereas those with $N_{eff}/L_{cov} > 100$ are considered as deep ('High' group). MSAs with $1 < N_{eff}/L_{cov} < 100$ are in the intermediate 'Medium' group. In our calculations, we consider the ratio between N_{eff} and the total number of positions L, which is equal to the length of the target sequence for both ProteinGym and ColabFold MSAs.

Generating the predictions with GEMME

GEMME takes as input a FASTA-formatted MSA, with the ungapped query sequence on top. We used the tool *reformat.pl* from the HH-suite [39] to convert A2M and A3M alignment files into FASTA format. Moreover, we modified the MSAs from ProteinNet and Pfam by putting the sequence of interest on top and removing the insertions with respect to this sequence. We used GEMME's Docker image, available from `http://www.lcqb.upmc.fr/GEMME`, to compute the predictions. For the proteins with only single mutations, we predicted the full mutational landscape with the command: "python2.7 $GEMME_PATH/gemme.py aliXXX.fasta -r input -f aliXXX.fasta" where *aliXXX.fasta* is the input MSA file in FASTA format. For the proteins with multiple mutations, we predicted only the effects of the mutations of interest. To do so, we passed a file specifying the list of mutations as input with the option "-m". We used the default parameters for all proteins and all input MSAs.

Assessing and comparing the predictions

Assessing and comparing the predictions obtained from the ProteinGym and ColabFold MSAs was straightforward since they cover the entire range of mutated positions and their query sequence is identical to the wild-type sequence used in the DMS. The MSAs from ProteinNet and Pfam however typically cover only a part of the mutated region and their query sequence sometimes display a few mutations with respect to

the DMS wild-type sequence. To compute the Spearman correlation, we restricted ourselves to the covered positions displaying the correct wild-type amino acid. When comparing two methods, we further reduced the calculation to their common positions.

Supplementary tables and figures

Table S1: **Coverage of the ProteinGym benchmark by the tested MSA generation protocols.** For each protein, we indicate its UniProt identifier, whether it is associated with measurements for multiple mutations, and whether the mutated region is covered by each of the tested protocols. We also give the PDB code selected for ProteinNet, and the number of Pfam domains (with available MSAs) overlapping with the mutated region.

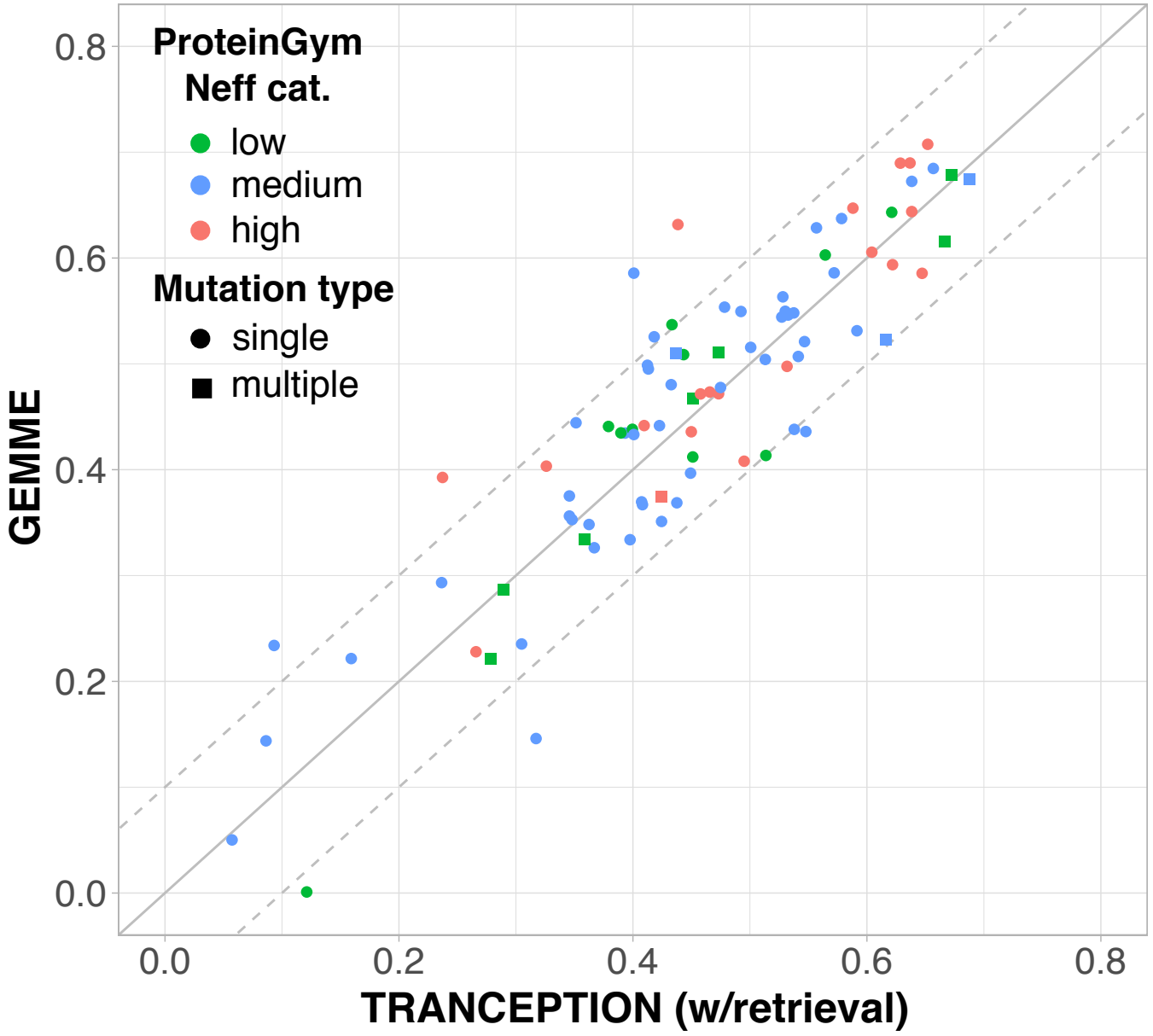

Figure S1: **Comparison of GEMME with the other mutational outcome predictor TRANCEPTION, given the same input MSAs.** Spearman rank correlation coefficients (ρ) are reported for the 87 DMS from the ProteinGym benchmark, using the ProteinGym MSAs as input (**Table 1**, see *ProteinGym*). The version of TRANCEPTION used here (with retrieval) combines a protein language model trained across families with information coming from a query-specific MSA retrieved at inference time [18]. The plotted values were taken from [18], where TRANCEPTION was shown to outperform Wavenet [23], DeepSequence [25], EVmutation [21], EVE [22], EMS-1v [27], and MSA Transformer [28]. GEMME predictions were generated using default parameters. The colors indicate the alignment depth categories defined in [18] (see also *Materials and Methods*). The shapes indicate whether the experiment contains only single mutations (circle) or also multiple mutations (square).

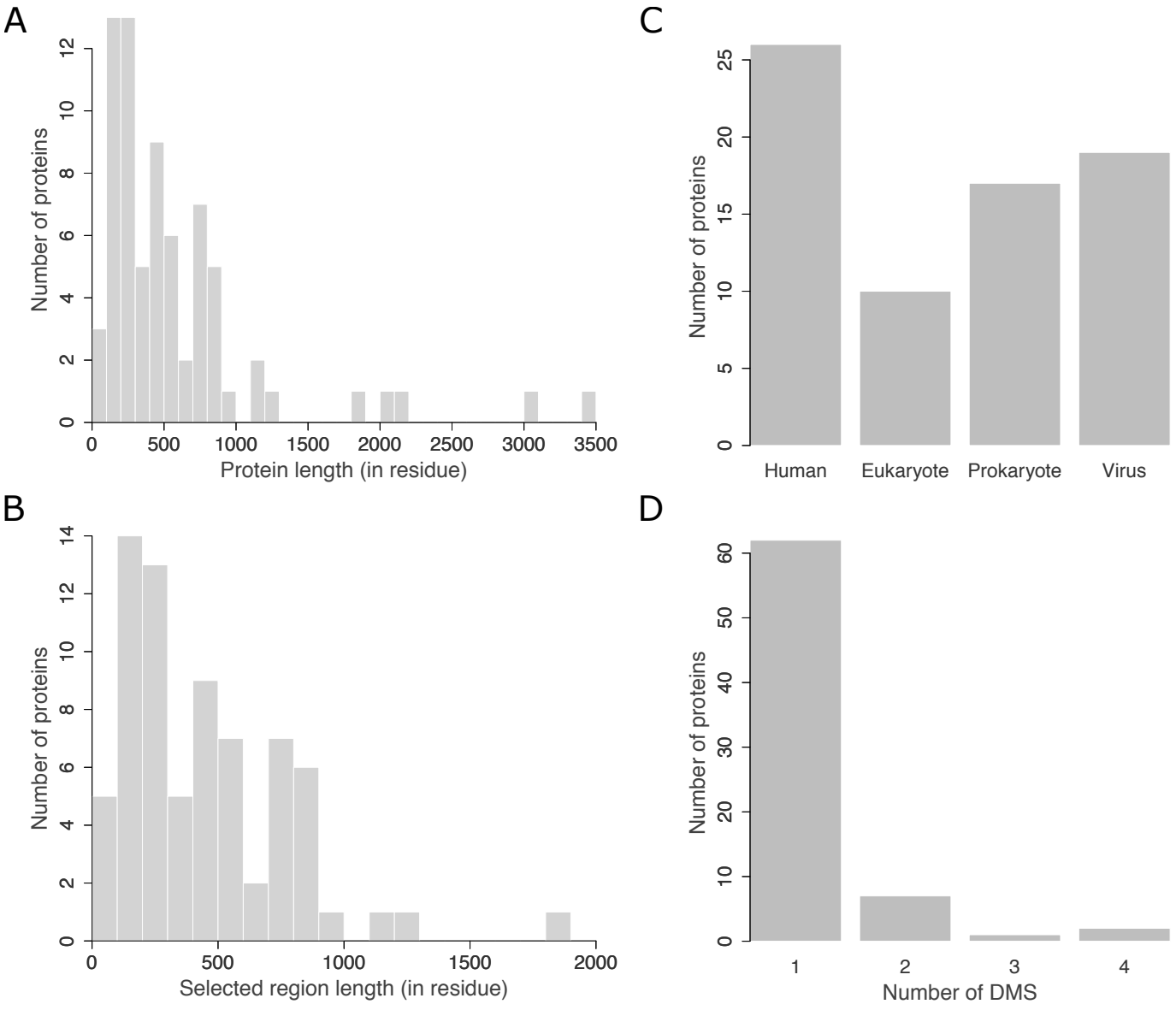

Figure S2: **ProteinGym benchmark properties. A.** Distribution of the length (in number of residues) of the 73 target protein sequences from the benchmark. **B.** Distribution of the length (in number of residues) of the protein regions covered by ProteinGym MSAs. **C.** Taxonomic classification of the proteins. The label "Eukaryote" refers to non-human eukaryotes. **D.** Distribution of the number of reported experiments per protein.

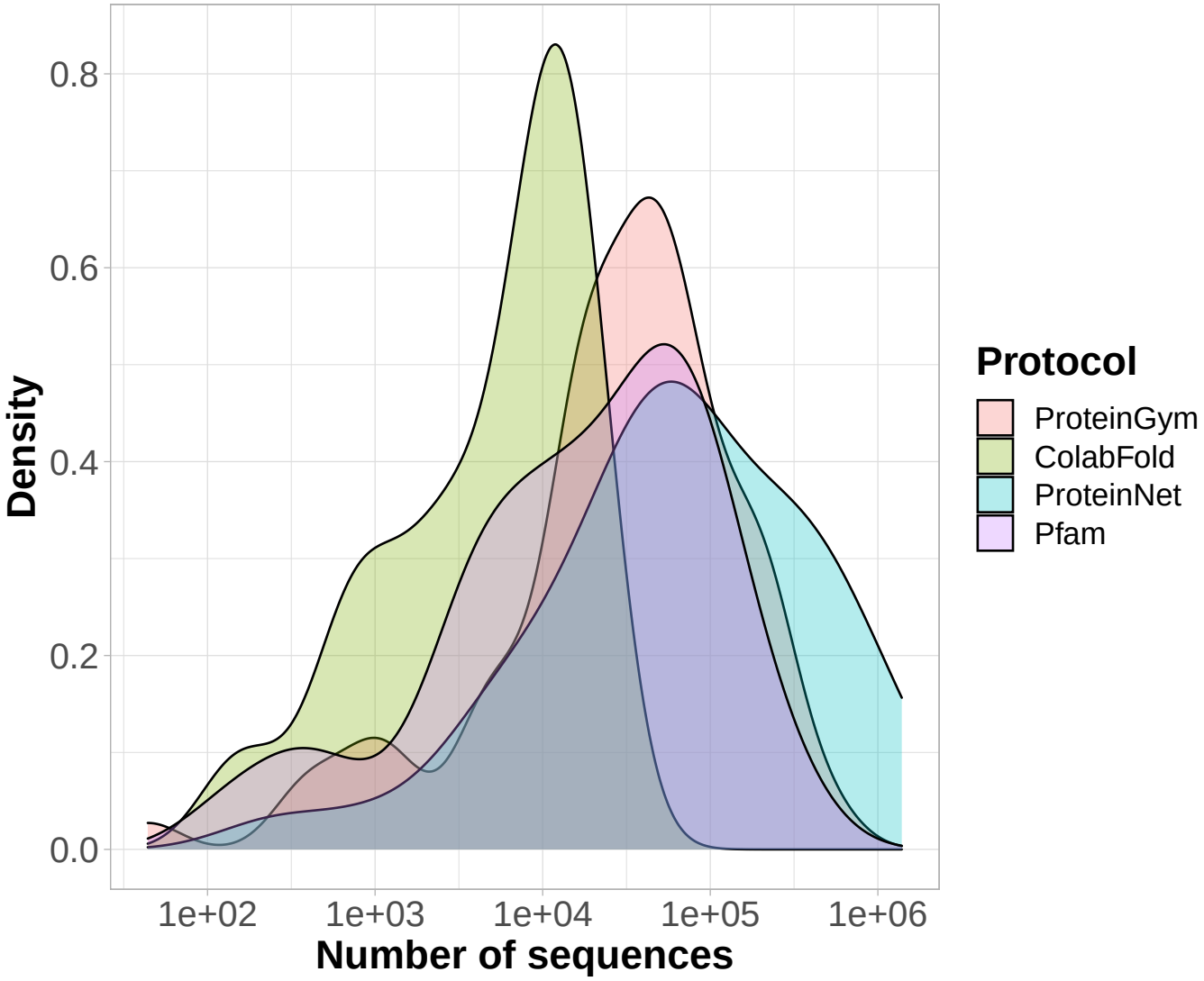

Figure S3: **Distribution of the number of sequences per MSA depending on the protocol.** The total number of MSAs varies from one protocol to another (see full details in **Table 1**).

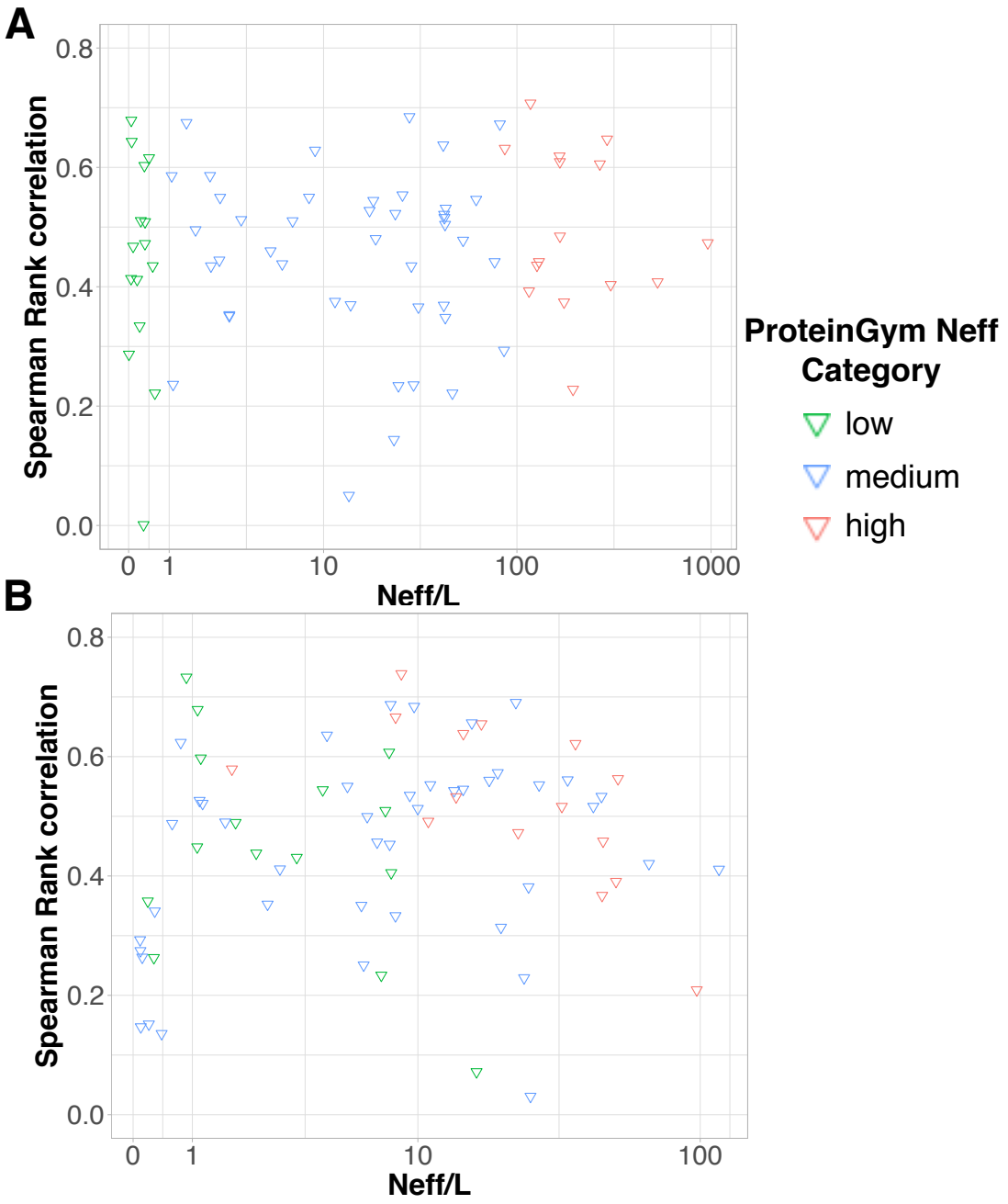

Figure S4: **Prediction accuracy in function of the alignment depth.** The input MSAs were generated using ProteinGym (**A**) or ColabFold (**B**) protocol. Each point (triangle) corresponds to a given input MSA (*i.e.* a given target sequence) and its y-value is averaged over the set of DMS experiments (between 1 and 4, see **Fig. S2**) associated to it. The Spearman correlations computed between the y (ρ) and log-x ($\log N_{eff}/L$) values are 0.065 and 0.225 for ProteinGym (**A**) and ColabFold (**B**), respectively. The colors indicate the ProteinGym N_{eff} categories, as defined in [18] (see also *Materials and Methods*). About half of the target sequences change category between the two protocols (see all red points, and also the blue points with a ratio lower than 1 and the green points with a ratio above 1 on panel B).

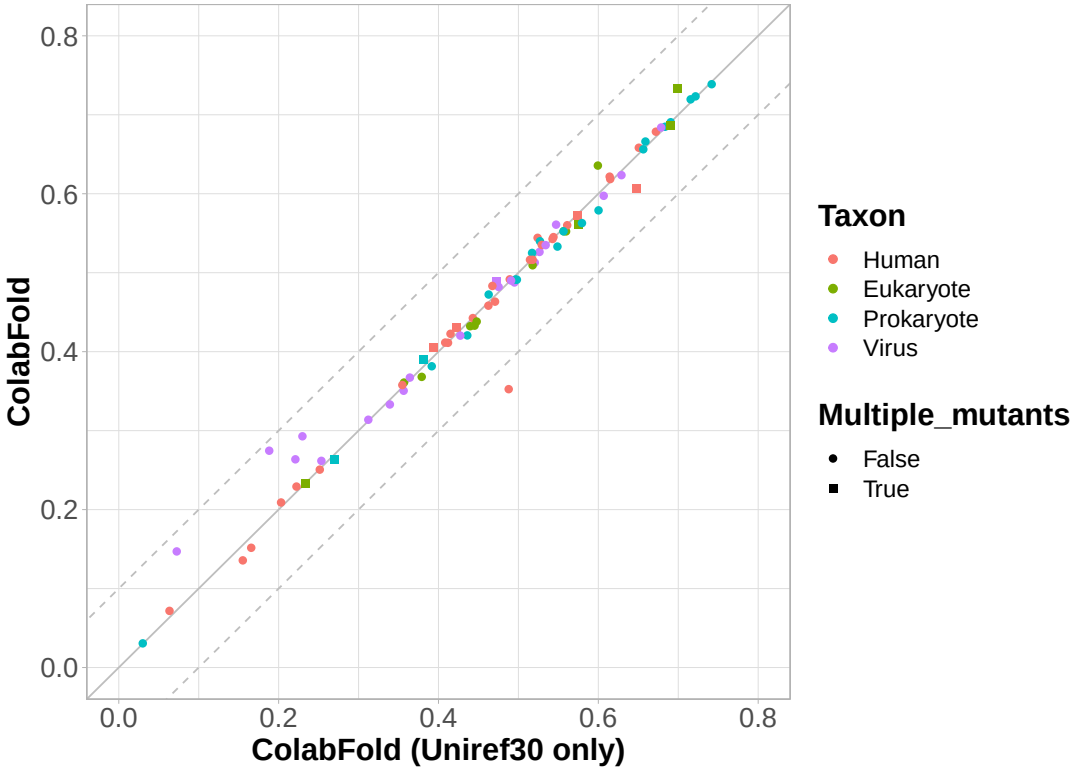

Figure S5: **Influence of the search database on GEMME performance.** The input MSAs were generated using the Colabfold protocol, considering only the UniRef30 database (x-axis) or both the UniRef30 database and the ColabFold database (y-axis). The values are reported for 86 out of the 87 DMS from ProteinGym. The DMS associated with SC6A4 is missing because the MSA generated from the UniRef30 database only was too shallow to compute reliable evolutionary conservation levels.

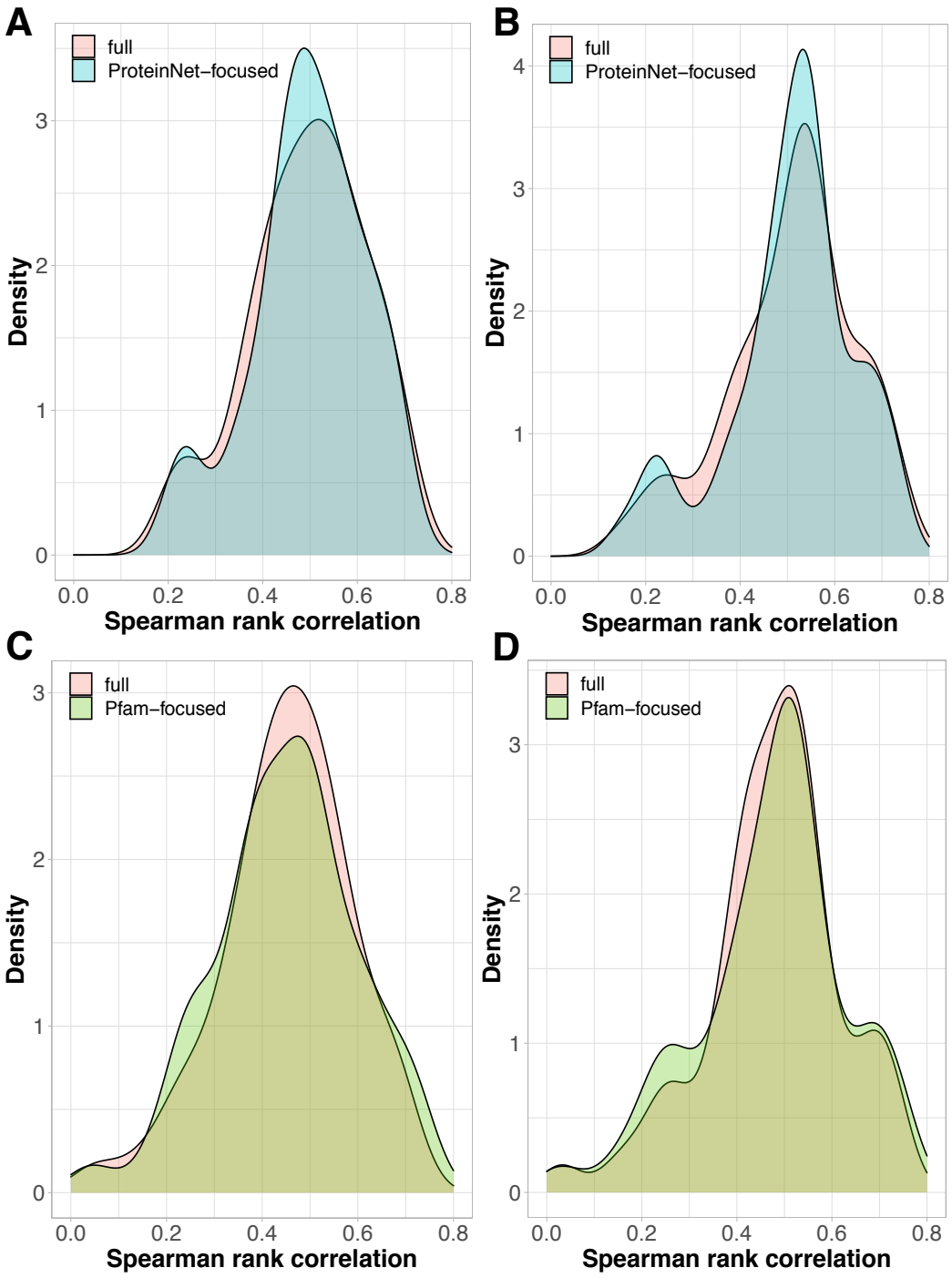

Figure S6: **Prediction accuracy achieved on the full-length versus partial proteins.** Distributions of Spearman rank correlations obtained with the ProteinGym (A,C) and ColabFold (B,D) protocols. **A-B.** Each distribution contains 51 values corresponding to the 51 DMS covered by ProteinNet. The correlations computed over the full-length proteins (in pink) are compared to those computed over the regions covered by ProteinNet (in blue). **C-D.** Each distribution contains 52 values corresponding to the 52 DMS covered by Pfam. The correlations computed over the full-length proteins (in pink) are compared to those computed over the regions covered by Pfam (in green).

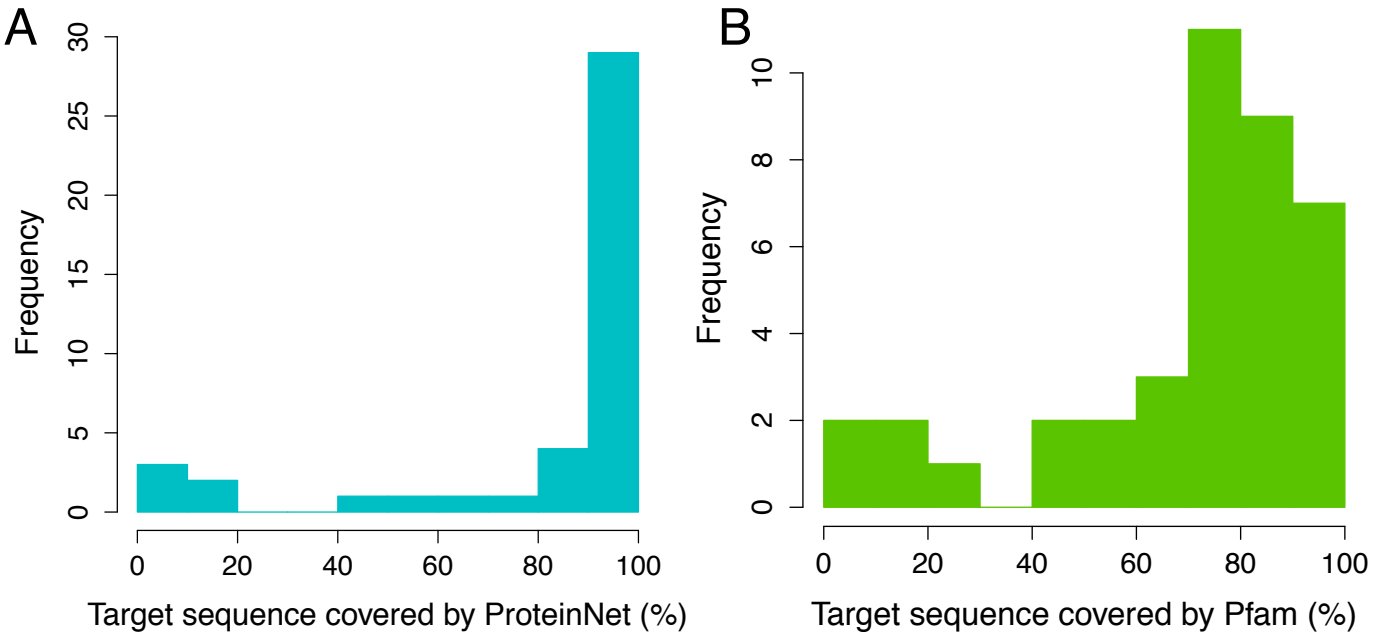

Figure S7: **Protein coverage from the ProteinNet and Pfam MSAs.** Percentage of residues from the target sequences covered by the MSAs from ProteinNet (**A**) and Pfam (**B**).